

# Randomized Frameproof Codes: Fingerprinting Plus Validation Minus Tracing

N. Prasanth Anthapadmanabhan  
Dept. of Electrical and Computer Eng.  
University of Maryland  
College Park, MD 20742  
Email: nagarajp@umd.edu

Alexander Barg  
Dept. of ECE and Inst. for Systems Research  
University of Maryland  
College Park, MD 20742  
Email: abarg@umd.edu

**Abstract**—We propose randomized frameproof codes for content protection, which arise by studying a variation of the Boneh-Shaw fingerprinting problem. In the modified system, whenever a user tries to access his fingerprinted copy, the fingerprint is submitted to a validation algorithm to verify that it is indeed permissible before the content can be executed. We show an improvement in the achievable rates compared to deterministic frameproof codes and traditional fingerprinting codes.

For coalitions of an arbitrary fixed size, we construct randomized frameproof codes which have an  $O(n^2)$  complexity validation algorithm and probability of error  $\exp(-\Omega(n))$ , where  $n$  denotes the length of the fingerprints. Finally, we present a connection between linear frameproof codes and minimal vectors for size-2 coalitions.

## I. INTRODUCTION

The availability of content (e.g., software, movies, music etc.) in the digital format, although with many advantages, has the downside that it is now easy for users to make copies, perform alterations, and share the content illegally. Thus there is a dire need for protecting the content against unauthorized redistribution, commonly termed as *piracy*.

In this paper, we consider a variation of the Boneh-Shaw fingerprinting scheme [6] for content protection. We start with an informal description of the problem. We will refer to the legal content owner as the *distributor* and the legitimate license holders as *users*. The distributor embeds a unique hidden mark, called a *fingerprint*, which identifies each licensed copy. The fingerprint locations, however, remain the same for all users. The collection of fingerprints is called the codebook and the distributor uses some form of randomization in choosing the codebook. We assume that changes to the actual content render it useless, while the fingerprint may be subject to alterations. This assumption is reasonable, for instance, in applications to software fingerprinting.

A single user is unable to pinpoint any of the fingerprint locations. However, if a set of users, called a *coalition of pirates*, compare their copies, they can infer some of the fingerprint locations by identifying the differences. The coalition now attempts to create a pirated copy with a forged fingerprint. In order to define the coalition's capability in creating the forgery, Boneh and Shaw introduced the *marking assumption*, which simply states that the coalition makes changes only in those positions where they find a difference (and hence are

definitely fingerprint locations) as they do not wish to damage the content permanently.

The objective of the distributor is to trace one of the guilty users whenever such a pirated copy is found. The maximum coalition size is a parameter of the problem. Such a collection of fingerprints together with the tracing algorithm is called a *fingerprinting code*. This problem has been studied in detail in [6], [4], [11], [2], where various constructions and upper bounds have been presented.

Consider now the modified system where each time a user accesses his fingerprinted copy, the fingerprint is validated to verify whether it is in fact permissible in the codebook being used and the execution continues only if the validation is successful. This limits the forgery possibilities for the pirates at the cost of an additional validation operation carried out every time a user accesses his copy. The idea is that by designing an efficient validation algorithm, we do not pay too high a price.

The advantage of this scheme is demonstrated by an improvement in the achievable rates compared to traditional fingerprinting codes, even though the actual property (cf. Definition 2.2) is not in general weaker than fingerprinting. In addition, since the pirates are limited to creating only a valid fingerprint and because we are interested in unique decoding, there is no additional tracing needed. The distributor simply accuses the user corresponding to the fingerprint in the pirated copy as guilty.

In this case, the coalition is successful if it is able to forge the fingerprint of an innocent user, thus “framing” him as the pirate. The distributor's objective is to design codes for which the probability that this error event occurs is small, deriving the name *frameproof codes*.

In the deterministic case with zero-error probability, frameproof codes arise as a special case of *separating codes*, which have been studied over many years since being introduced in [8]. For further references on deterministic frameproof codes and separating codes, we refer the interested reader to [9], [7], [10], [5]. In order to emphasize the difference that we consider the randomized setting, we call our codes *randomized frameproof codes*.

The rest of the paper is organized as follows. In Section II, we give a formal definition for randomized frameproof codes. Achievable rates under no restrictions on validation

complexity are presented in Section III. In Section IV, we show the existence of linear frameproof codes and exhibit a connection to minimal vectors for size-2 coalitions. Finally, we design a concatenated code with efficient validation for arbitrary coalition sizes in Section V.

## II. PROBLEM DEFINITION

We will use the following notation. Boldface will denote vectors. The Hamming distance between vectors  $\mathbf{x}_1, \mathbf{x}_2$  will be denoted by  $\text{dist}(\mathbf{x}_1, \mathbf{x}_2)$ . We also write  $s_z(\mathbf{x}_1, \dots, \mathbf{x}_t)$  to denote the number of  $z^T$  columns in the matrix formed with the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_t$  as the rows. For a positive integer  $n$ , the shorthand notation  $[n]$  will stand for the set  $\{1, \dots, n\}$ . We use  $h(p) := -p \log_2 p - (1-p) \log_2 (1-p)$  to denote the binary entropy function and  $D(p||q) := p \log_2 (p/q) + (1-p) \log_2 ((1-p)/(1-q))$  to denote the information divergence.

Let  $\mathcal{Q}$  be an alphabet (often a field) of finite size  $q$  and let  $M$  be the number of users in the system. Assume that there is some ordering of the users and denote their set by  $[M]$ . The fingerprint for each user is of length  $n$ .

Consider the following random experiment. We have a family of  $q$ -ary codes  $\{C_k, k \in \mathcal{K}\}$  of length  $n$  and size  $M$ . In particular, here the code  $C_k$  refers to an *ordered set* of  $M$  codewords. We pick one of the codes according to the probability distribution function  $(\pi(k), k \in \mathcal{K})$ . For brevity, the result of this random experiment is called a *randomized code* and is denoted by  $\mathcal{C}$ . The *rate* of this code is  $R = n^{-1} \log_q M$ . We will refer to elements of the set  $\mathcal{K}$  as *keys*. Note that the dependence on  $n$  has been suppressed for simplicity.

The distributor assigns the fingerprints as follows. He chooses one of the keys, say  $k$ , with probability  $\pi(k)$ , and assigns to user  $i$  the  $i$ th codeword of  $C_k$ , denoted by  $C_k(i)$ . Following the standard cryptographic precept that the adversary knows the system, we allow the users to be aware of the family of codes  $\{C_k\}$  and the distribution  $\pi(\cdot)$ , but the exact key choice is kept secret by the distributor.

The fingerprints are assumed to be distributed within the host message in some fixed locations unknown to the users. Before a user executes his copy, his fingerprint is submitted to a *validation* algorithm, which checks whether the fingerprint is a valid codeword in the current codebook. The execution continues only if the validation succeeds.

A *coalition*  $U$  of  $t$  users is an arbitrary  $t$ -subset of  $[M]$ . The members of the coalition are commonly referred to as *pirates*. Suppose the collection of fingerprints assigned to  $U$ , namely  $C_k(U)$ , is  $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ . The goal of the pirates is to create a forged fingerprint different from theirs which is valid under the current key choice.

Coordinate  $i$  of the fingerprints is called *undetectable* for the coalition  $U$  if  $x_{1i} = x_{2i} = \dots = x_{ti}$  and is called *detectable* otherwise. We assume that the coalition follows the *marking assumption* [6] in creating the forgery.

**Definition 2.1:** The *marking assumption* states that for any fingerprint  $\mathbf{y}$  created by the coalition  $U$ ,  $y_i = x_{1i} = x_{2i} = \dots = x_{ti}$  in every coordinate  $i$  that is undetectable.

In other words, in creating  $\mathbf{y}$ , the pirates can modify only detectable positions.

For a given set of observed fingerprints  $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ , the set of forgeries that can be created by the coalition is called the *envelope*. Its definition depends on the exact rule the coalition should follow to modify the detectable positions [4]:

- If the coalition is restricted to use only a symbol from their assigned fingerprints in the detectable positions, we obtain the *narrow-sense envelope*:

$$e(\mathbf{x}_1, \dots, \mathbf{x}_t) = \{\mathbf{y} \in \mathcal{Q}^n | y_i \in \{x_{1i}, \dots, x_{ti}\}, \forall i \in [n]\}; \quad (1)$$

- If the coalition can use any symbol from the alphabet in the detectable positions, we obtain the *wide-sense envelope*:

$$E(\mathbf{x}_1, \dots, \mathbf{x}_t) = \{\mathbf{y} \in \mathcal{Q}^n | y_i = x_{1i}, \forall i \text{ undetectable}\}. \quad (2)$$

For the binary alphabet, both envelopes are exactly the same. In the following, we will use  $\mathcal{E}(\cdot)$  to denote the envelope from any of the rules mentioned above.

**Definition 2.2:** A randomized code  $\mathcal{C}$  is said to be *t-frameproof* with  $\varepsilon$ -error if for all  $U \subseteq [M]$  such that  $|U| \leq t$ , it holds that

$$\Pr\{\mathcal{E}(\mathcal{C}(U)) \cap (\mathcal{C} \setminus \mathcal{C}(U)) \neq \emptyset\} \leq \varepsilon, \quad (3)$$

where the probability is taken over the distribution  $\pi(\cdot)$ .

**Remark 2.3:** Note that the *t-frameproof* property as defined above is not in general weaker than the *t-fingerprinting* property, i.e., a code which is *t-fingerprinting* with  $\varepsilon$ -error [6, Definition IV.2] is not automatically *t-frameproof* with  $\varepsilon'$ -error, for any  $0 \leq \varepsilon' < 1$ .

A straightforward extension of the fingerprinting definition yields a randomized code which satisfies the following condition: For any coalition of size at most  $t$  and any strategy they may use in devising a forgery, the probability that the forgery is valid is small. However, this definition would trivially enable us to achieve arbitrarily high rates. Hence, we use the above (stronger) definition.

## III. LOWER BOUNDS FOR BINARY FRAMEPROOF CODES

Let us construct a binary randomized code  $\mathcal{C}$  of length  $n$  and size  $M = 2^{nR}$  as follows. We pick each entry in the  $M \times n$  matrix independently to be 1 with probability  $p$ , for some  $0 \leq p \leq 1$ .

**Theorem 3.1:** The randomized code  $\mathcal{C}$  is *t-frameproof* with error probability decaying exponentially in  $n$  for any rate

$$R < -p^t \log_2 p - (1-p)^t \log_2 (1-p). \quad (4)$$

**Proof:** For  $\gamma > 0$ , define the set of  $t$ -tuples of vectors

$$\mathcal{T}_{t,\gamma} := \left\{ (\mathbf{x}_1, \dots, \mathbf{x}_t) : \begin{array}{l} s_1(\mathbf{x}_1, \dots, \mathbf{x}_t) \in I_\gamma, \\ s_0(\mathbf{x}_1, \dots, \mathbf{x}_t) \in J_\gamma \end{array} \right\},$$

where  $I_\gamma := [n(p^t - \gamma), n(p^t + \gamma)]$  and  $J_\gamma := [n((1-p)^t - \gamma), n((1-p)^t + \gamma)]$ . It is clear that for any coalition  $U$  of size  $t$ , the observed fingerprints  $(\mathbf{x}_1, \dots, \mathbf{x}_t)$  belong to  $\mathcal{T}_{t,\gamma}$  with

high probability<sup>1</sup>. Hence, we will refer to  $\mathcal{T}_{t,\gamma}$  as the set of *typical* fingerprints. For any coalition  $U$  of size  $t$

$$\begin{aligned} & \Pr\{\mathcal{E}(\mathcal{C}(U)) \cap (\mathcal{C} \setminus \mathcal{C}(U)) \neq \emptyset\} \\ & \leq \Pr\{\mathcal{C}(U) \notin \mathcal{T}_{t,\gamma}\} \\ & \quad + \Pr\{\exists \mathbf{y} \in \mathcal{C} \setminus \mathcal{C}(U) : \mathbf{y} \in \mathcal{E}(\mathcal{C}(U)) | \mathcal{C}(U) \in \mathcal{T}_{t,\gamma}\}. \end{aligned} \quad (5)$$

The first term in the above equation decays exponentially in  $n$ . It is left to prove that the second term is also exponentially decaying for  $R$  satisfying (4).

A codeword in  $\mathcal{C} \setminus \mathcal{C}(U)$  is a part of  $\mathcal{E}(\mathcal{C}(U))$  if it contains a 1 (resp. 0) in all  $s_1(\mathcal{C}(U))$  (resp.  $s_0(\mathcal{C}(U))$ ) positions. Since  $\mathcal{C}(U) \in \mathcal{T}_{t,\gamma}$ , by taking a union bound the second term in (5) is at most

$$2^{nR} p^{n(p^t - \gamma)} (1 - p)^{n((1-p)^t - \gamma)},$$

which decays exponentially in  $n$  for

$$R < -(p^t - \gamma) \log_2 p - ((1 - p)^t - \gamma) \log_2 (1 - p).$$

The proof is completed by taking  $\gamma$  to be arbitrarily small. ■

The bias  $p$  in the construction of  $\mathcal{C}$  can be chosen optimally for each value of  $t$ . Numerical values of the rate thus obtained are shown in Table I, where they are compared with the existence bounds for deterministic zero-error frameproof codes (from [7]) and rates of fingerprinting codes (from [2], [1]). Observe that there is a factor of  $t$  improvement compared to the rate of deterministic frameproof codes.

TABLE I  
COMPARISON OF RATES

$t$	Rates		
	Randomized Frameproof	Deterministic Frameproof	Fingerprinting
2	0.5	0.2075	0.25
3	0.25	0.0693	0.0833
4	0.1392	0.04	0.0158
5	0.1066	0.026	0.0006

#### IV. LINEAR FRAMEPROOF CODES

Unlike fingerprinting codes, randomized frameproof codes eliminate the need for a tracing algorithm, but the fingerprints still need to be validated. As the validation algorithm is executed everytime a user accesses his copy, we require that this algorithm have an efficient running time. Although the codes designed in the previous section have high rates, they come at the price of an  $\exp(n)$  complexity validation algorithm. Linear codes are an obvious first choice in trying to design efficient frameproof codes as they can be validated in  $O(n^2)$  time by simply verifying the parity-check equations.

<sup>1</sup>We say that an event occurs with high probability if the probability that it fails is at most  $\exp(-cn)$ , where  $c$  is a positive constant.

##### A. Linear construction for $t = 2$

We now present a binary linear frameproof code for  $t = 2$  which achieves the rate given by Theorem 3.1. Suppose we have  $M = 2^{nR}$  users. We construct a randomized linear code  $\mathcal{C}$  as follows. Pick a random  $n(1 - R) \times n$  parity-check matrix with each entry chosen independently to be 0 or 1 with equal probability. The corresponding set of binary vectors which satisfy the parity-check matrix form a linear code of size  $2^{nR}$  with high probability. Each user is then assigned a unique codeword selected uniformly at random from this collection. In the few cases that the code size exceeds  $2^{nR}$ , we simply ignore the remaining codewords during the assignment. However, note that since the validation algorithm simply verifies the parity-check equations, it will pronounce the ignored vectors also as valid.

*Theorem 4.1:* The randomized linear code  $\mathcal{C}$  is 2-frameproof with error probability decaying exponentially in  $n$  for any rate  $R < 0.5$ .

*Proof:* As in the proof of Theorem 3.1, we begin by defining the set of typical pairs of fingerprints. For  $\gamma > 0$ , define

$$\mathcal{T}_\gamma := \left\{ (\mathbf{x}_1, \mathbf{x}_2) : s_{ij}(\mathbf{x}_1, \mathbf{x}_2) \in I_\gamma, \forall i, j \in \{0, 1\} \right\},$$

where  $I_\gamma := [n(1/4 - \gamma), n(1/4 + \gamma)]$ . For any coalition  $U$  of two users

$$\begin{aligned} & \Pr\{\mathcal{E}(\mathcal{C}(U)) \cap (\mathcal{C} \setminus \mathcal{C}(U)) \neq \emptyset\} \\ & \leq \Pr\{\mathcal{C}(U) \notin \mathcal{T}_\gamma\} + \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{T}_\gamma} \Pr\{\mathcal{C}(U) = (\mathbf{x}_1, \mathbf{x}_2)\} \\ & \quad \times \Pr\{\exists \mathbf{y} \in \mathcal{C} : \mathbf{y} \in \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2) \setminus \{\mathbf{x}_1, \mathbf{x}_2\} | \mathcal{C}(U) = (\mathbf{x}_1, \mathbf{x}_2)\}. \end{aligned}$$

It can be seen that the first term again decays exponentially in  $n$ . We now consider the term inside the summation

$$\Pr\{\exists \mathbf{y} \in \mathcal{C} : \mathbf{y} \in \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2) \setminus \{\mathbf{x}_1, \mathbf{x}_2\} | \mathcal{C}(U) = (\mathbf{x}_1, \mathbf{x}_2)\}.$$

Observe that for any two binary vectors  $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{T}_\gamma$ , the sum  $\mathbf{x}_1 + \mathbf{x}_2 \notin \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2)$  and also  $\mathbf{0} \notin \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2)$ . Therefore, every vector in  $\mathcal{E}(\mathbf{x}_1, \mathbf{x}_2) \setminus \{\mathbf{x}_1, \mathbf{x}_2\}$  is linearly independent from  $\mathbf{x}_1, \mathbf{x}_2$ . Thus for any  $\mathbf{y} \in \mathcal{E}(\mathbf{x}_1, \mathbf{x}_2) \setminus \{\mathbf{x}_1, \mathbf{x}_2\}$ ,

$$\Pr\{\mathbf{y} \in \mathcal{C} | \mathcal{C}(U) = (\mathbf{x}_1, \mathbf{x}_2)\} = \Pr\{\mathbf{y} \in \mathcal{C}\} = 2^{-n(1-R)}.$$

Since  $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{T}_\gamma$ ,  $|\mathcal{E}(\mathbf{x}_1, \mathbf{x}_2)| \leq 2^{n(1/2 + 2\gamma)}$ . By taking the union bound and  $\gamma$  to be arbitrarily small, we obtain the result. ■

##### B. Connection to minimal vectors

In this subsection, we show a connection between linear 2-frameproof codes and minimal vectors. We first recall the definition for minimal vectors (see, for e.g., [3]). Let  $C$  be a  $q$ -ary  $[n, k]$  linear code. The support of a vector  $\mathbf{c} \in C$  is given by  $\text{supp}(\mathbf{c}) = \{i \in [n] : c_i \neq 0\}$ . We write  $\mathbf{c}' \preceq \mathbf{c}$  if  $\text{supp}(\mathbf{c}') \subseteq \text{supp}(\mathbf{c})$ .

*Definition 4.2:* A nonzero vector  $\mathbf{c} \in C$  is called *minimal* if  $\mathbf{0} \neq \mathbf{c}' \preceq \mathbf{c}$  implies  $\mathbf{c}' = \alpha \mathbf{c}$ , where  $\mathbf{c}'$  is another code vector and  $\alpha$  is a nonzero constant.

*Proposition 4.3:* For any  $\mathbf{x}_1, \mathbf{x}_2 \in C$ ,  $\mathbf{x}_1 \neq \mathbf{x}_2$ , if  $\mathbf{x}_2 - \mathbf{x}_1$  is minimal then  $e(\mathbf{x}_1, \mathbf{x}_2) \cap (C \setminus \{\mathbf{x}_1, \mathbf{x}_2\}) = \emptyset$ . If  $q = 2$ , the converse is also true.

*Proof:* Consider any  $\mathbf{y} \in \mathcal{Q}^n$  and define the translate  $\mathbf{y}' := \mathbf{y} - \mathbf{x}_1$ . It follows that

$$\mathbf{y} \in C \Leftrightarrow \mathbf{y}' \in C \quad (6)$$

$$\mathbf{y} \notin \{\mathbf{x}_1, \mathbf{x}_2\} \Leftrightarrow \mathbf{y}' \notin \{\mathbf{0}, \mathbf{x}_2 - \mathbf{x}_1\}. \quad (7)$$

Furthermore, if  $y_i \in \{x_{1i}, x_{2i}\}$ , then  $y'_i \in \{0, x_{2i} - x_{1i}\}$  for all  $i \in [n]$ . Therefore,

$$\mathbf{y} \in e(\mathbf{x}_1, \mathbf{x}_2) \Rightarrow \begin{cases} \mathbf{y}' \preceq \mathbf{x}_2 - \mathbf{x}_1, \\ \mathbf{y}' \neq \alpha(\mathbf{x}_2 - \mathbf{x}_1), \forall \alpha \notin \{0, 1\}. \end{cases} \quad (8)$$

Using (6), (7), (8), we obtain that  $e(\mathbf{x}_1, \mathbf{x}_2) \cap (C \setminus \{\mathbf{x}_1, \mathbf{x}_2\}) \neq \emptyset$  implies that  $\mathbf{x}_2 - \mathbf{x}_1$  is non-minimal.

For  $q = 2$ , it is easily seen that the reverse statement also holds in (8) and thus the converse is also true. ■

Recall the random linear code constructed by generating a random  $n(1 - R) \times n$  parity-check matrix in the previous subsection. With some abuse of notation, let us denote the (unordered) set of vectors satisfying the random parity-check matrix also by  $C$ . Let  $\mathcal{M}(C)$  denote the set of minimal vectors in  $C$ . We have the following companion result to Corollary 2.5 in [3].

*Corollary 4.4:* As  $n \rightarrow \infty$ ,

$$\mathbb{E} \left[ \frac{|\mathcal{M}(C)|}{|C|} \right] = \begin{cases} 1, & R < 1/2 \\ 0, & R > 1/2 \end{cases}$$

*Proof:* As a consequence of Proposition 4.3, for any two users  $\{u_1, u_2\}$ , we obtain

$$\begin{aligned} & \Pr\{\mathcal{E}(C(u_1, u_2)) \cap (C \setminus \mathcal{C}(u_1, u_2)) \neq \emptyset\} \\ &= \Pr\{C(u_2) - C(u_1) \notin \mathcal{M}(C)\} \\ &= 1 - \mathbb{E} \left[ \frac{|\mathcal{M}(C)|}{|C| - 1} \right]. \end{aligned}$$

The first part of the result is now true by Theorem 4.1. We skip the details of the latter part which is easily proved using Chernoff bounds. ■

### C. Linear codes for larger $t$

In the light of Theorem 4.1, a natural question to ask is whether there exist randomized linear frameproof codes for  $t > 2$ , perhaps allowing even a larger alphabet. It turns out that, just as in the deterministic case, linear frameproof codes do not always exist in the randomized setting too.

*Proposition 4.5:* There do not exist  $q$ -ary linear  $t$ -frameproof codes with  $\varepsilon$ -error,  $0 \leq \varepsilon < 1$ , which are secure with the wide-sense envelope if either  $t > q$  or  $q > 2$ .

*Proof:* Consider a coalition of  $q + 1$  users. For any linear code realized from the family where the observed fingerprints are, say,  $\mathbf{x}_1, \dots, \mathbf{x}_{q+1}$ , the forgery  $\mathbf{y} = \mathbf{x}_1 + \dots + \mathbf{x}_{q+1}$  is a part of  $E(\mathbf{x}_1, \dots, \mathbf{x}_{q+1})$ . In addition, it is also a valid fingerprint as the code is linear. This proves the first part of the proposition.

To prove the second part, consider an alphabet (a field) with  $q > 2$ . For any two pirates with fingerprints  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , the

forgery  $\mathbf{y} = \alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2$ , where  $\alpha \neq 0, 1$ , is a valid codeword (by linearity) and is also a part of the wide-sense envelope. ■

Consequently, in considering linear frameproof codes which are wide-sense secure, we are limited to  $t = 2, q = 2$ .

### V. POLYNOMIAL-TIME VALIDATION FOR LARGER $t$

Usually, the amount of redundancy needed increases with the alphabet size in fingerprinting applications. Thus, we are mainly interested in constructing *binary* frameproof codes which have polynomial-time validation. With the binary alphabet, there is no distinction between wide-sense and narrow-sense envelopes. Therefore, there do not exist binary linear frameproof codes for  $t > 2$  by Proposition 4.5. In this section, we use the idea of code concatenation to construct a binary frameproof code with polynomial-time validation.

In the case of deterministic codes, if both the inner and outer codes are  $t$ -frameproof ( $(t, 1)$ -separating) with zero-error, then the concatenated code is also  $t$ -frameproof. We will now establish a similar result when the inner code is a randomized  $t$ -frameproof code.

Let the outer code  $C_{\text{out}}$  be a (deterministic)  $q$ -ary linear  $[N, K, \Delta]$  code. For each of the  $N$  coordinates of the outer code, generate an independent instance of a randomized binary code  $C_{\text{in}}$  of length  $m$  and size  $q$  which is  $t$ -frameproof with  $\varepsilon$ -error. Then the concatenated code  $C$  with outer code  $C_{\text{out}}$  and inner code independent instances of  $C_{\text{in}}$  is a randomized binary code of length  $n = Nm$  and size  $q^K$ .

*Theorem 5.1:* If the relative minimum distance of  $C_{\text{out}}$  satisfies

$$\frac{\Delta}{N} \geq 1 - \frac{1}{t}(1 - \xi) \quad (9)$$

and the error probability  $\varepsilon < \xi$  for  $C_{\text{in}}$ , then the concatenated code  $C$  is  $t$ -frameproof with error probability  $2^{-ND(\xi|\varepsilon)}$  and has a  $\text{poly}(n)$  validation algorithm.

*Proof:* In the proof, all vectors are  $q$ -ary corresponding to the outer alphabet. Define

$$\begin{aligned} s(\mathbf{y}, \{\mathbf{x}_1, \dots, \mathbf{x}_t\}) &:= |\{i \in [N] : y_i \in \{x_{1i}, \dots, x_{ti}\}\}|, \\ d(\mathbf{y}, \{\mathbf{x}_1, \dots, \mathbf{x}_t\}) &:= \min_{i \in [t]} \text{dist}(\mathbf{y}, \mathbf{x}_i). \end{aligned}$$

Consider a coalition  $U \subseteq \{1, \dots, q^K\}$  of size  $t$ . For any coordinate  $i \in [N]$  of the outer code, the coalition observes at most  $t$  different symbols of the outer alphabet, i.e., at most  $t$  different codewords of the inner code. Thus if the  $t$ -frameproof property holds for the observed symbols for the realization of  $C_{\text{in}}$  at coordinate  $i$ , then at the outer level the coalition is restricted to output one of the symbols it observes, i.e., the narrow-sense rule (1) holds. On the other hand, a failure of the  $t$ -frameproof property at the inner level code implies that the coalition is able to create a symbol different from what they observe in the corresponding coordinate at the outer level.

Accordingly, let  $\chi_i, i = 1, \dots, N$ , denote the indicator random variables (r.v.s) for failures at the inner level with  $\Pr\{\chi_i = 1\} \leq \varepsilon$  since the inner code has  $\varepsilon$ -error. Note that  $\chi_i$  are independent because we have an independent

instance of the randomized code for every  $i = 1, \dots, N$ . Then  $Z = \sum_{i=1}^N \chi_i$  is a Binomial r.v. denoting the number of coordinates where the narrow-sense rule fails at the outer level. For  $0 \leq z \leq N$ , let  $e_z(\cdot)$  denote the envelope when the narrow-sense rule is followed only in some  $N - z$  outer-level coordinates, i.e.,

$$e_z(\mathbf{x}_1, \dots, \mathbf{x}_t) = \{\mathbf{y} : s(\mathbf{y}, \{\mathbf{x}_1, \dots, \mathbf{x}_t\}) \geq N - z\}.$$

For any  $\mathbf{y} \in e_z(\mathbf{x}_1, \dots, \mathbf{x}_t)$ , there exists some  $l \in \{1, \dots, t\}$  such that  $s(\mathbf{y}, \mathbf{x}_l) \geq (N - z)/t$ , i.e.,  $\text{dist}(\mathbf{y}, \mathbf{x}_l) \leq N - (N - z)/t$ . Therefore,

$$e_z(\mathbf{x}_1, \dots, \mathbf{x}_t) \subseteq \left\{ \mathbf{y} : d(\mathbf{y}, \{\mathbf{x}_1, \dots, \mathbf{x}_t\}) \leq N - \frac{N - z}{t} \right\}. \quad (10)$$

The coalition  $U$  succeeds when it creates a forgery which is valid in the outer code. Thus the probability of error is at most

$$\begin{aligned} & \Pr\{\exists \mathbf{y} \in C_{\text{out}} \setminus C_{\text{out}}(U) : \mathbf{y} \in e_Z(C_{\text{out}}(U))\} \\ & \leq \Pr\left\{ \exists \mathbf{y} \in C_{\text{out}} \setminus C_{\text{out}}(U) : d(\mathbf{y}, C_{\text{out}}(U)) \leq N - \frac{N - Z}{t} \right\} \end{aligned} \quad (11)$$

$$= \Pr\left\{ N - \frac{N - Z}{t} \geq \Delta \right\} \quad (12)$$

$$\leq \Pr\{Z \geq N\xi\} \quad (13)$$

$$\leq 2^{-ND(\xi||\varepsilon)}, \quad (14)$$

where (11) follows from (10), (12) is because  $C_{\text{out}}$  is a linear code with minimum distance  $\Delta$ , (13) is due to the condition (9) and (14) is obtained by standard large deviation bounds.

The validation algorithm operates in two steps. In the first step, the inner code is decoded/validated for every outer code coordinate by exhaustive search over  $q$  codewords. We then check whether the resulting  $q$ -ary vector is a member of the outer code by verifying the parity-check equations. The claim about the polynomial-time complexity is true by choosing an appropriate scaling for the inner code length, for instance,  $m \sim \log_2(n)$ . ■

We now make specific choices for the outer and inner codes in Theorem 5.1 to arrive at explicit constructions. We take  $C_{\text{in}}$  to be the binary randomized  $t$ -frameproof code presented in Theorem 3.1 and with growing length. Thus we have the inner code rate as

$$R_t = \max_{p \in [0,1]} [-p^t \log_2 p - (1 - p)^t \log_2 (1 - p)]$$

and error probability  $\varepsilon = 2^{-m\beta}$  for some  $\beta > 0$ . The outer code  $C_{\text{out}}$  is a  $[q - 1, K]$  Reed-Solomon (RS) code with rate at most  $(1 - \xi)/t$  in order to satisfy the condition (9) on the minimum distance. Observe that for  $\varepsilon$  approaching 0 (for large  $m$ ) and  $\xi$  fixed,  $D(\xi||\varepsilon) \sim \xi \log_2(1/\varepsilon)$ . Therefore, with  $\varepsilon = 2^{-m\beta}$ , the error probability of the concatenated code is at most  $2^{-n(\xi\beta + o(1))}$ . By taking  $\xi$  arbitrarily small and  $m$  sufficiently large to satisfy  $\varepsilon < \xi$ , we obtain the following result.

**Corollary 5.2:** The binary randomized code obtained by concatenating  $C_{\text{out}}$  and  $C_{\text{in}}$  is  $t$ -frameproof with error prob-

ability  $\exp(-\Omega(n))$ , validation complexity  $O(n^2)$  and rate arbitrarily close to  $R_t/t$ .

## VI. CONCLUSION

The question of upper bounds on the rate of randomized frameproof codes is open.

## ACKNOWLEDGMENTS

The research is supported in part by NSF grants CCF0515124 and CCF0635271, and by NSA grant H98230-06-1-0044.

## REFERENCES

- [1] N. P. Anthapadmanabhan and A. Barg, "Random binary fingerprinting codes for arbitrarily sized coalitions," *Proc. IEEE Internat. Sympos. Inform. Theory (ISIT 2006)*, pp. 351-355, 2006.
- [2] N. P. Anthapadmanabhan, A. Barg and I. Dumer, "On the fingerprinting capacity under the marking assumption," *IEEE Trans. on Inform. Theory - Special Issue on Information-Theoretic Security*, Jun. 2008, to appear. Available at <http://arxiv.org/abs/cs.IT/0612073>.
- [3] A. Ashikhmin and A. Barg, "Minimal vectors in linear codes," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 2010-2017, Sep. 1998.
- [4] A. Barg, G. R. Blakley and G. Kabatiansky, "Digital fingerprinting codes: Problem statements, constructions, identification of traitors," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 852-865, Apr. 2003.
- [5] S. R. Blackburn, "Combinatorial schemes for protecting digital content," *Surveys in combinatorics*, 2003 (Bangor), London Math. Soc. Lecture Note Ser., vol. 307, pp. 43-78, Cambridge Univ. Press, Cambridge, 2003.
- [6] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1897-1905, Sep. 1998.
- [7] G. Cohen and H. G. Schaathun, "Asymptotic overview of separating codes," Report no. 248, Department of Informatics, University of Bergen, 52pp., May 2003. Available at [www.ii.uib.no](http://www.ii.uib.no).
- [8] A. D. Friedman, R. L. Graham, and J. D. Ullman, "Universal single transition time asynchronous state assignments," *IEEE Trans. Comput.*, vol. C-18, pp. 541-547, 1969.
- [9] Y. L. Sagalovich, "Separating systems," *Probl. Inform. Trans.*, vol. 30, no. 2, pp. 105-123, 1994.
- [10] J. N. Staddon, D. R. Stinson and R. Wei, "Combinatorial properties of frameproof and traceability codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 1042-1049, Mar. 2001.
- [11] G. Tardos, "Optimal probabilistic fingerprint codes," *Journal of the ACM*, to appear. Preliminary version in *Proc. 35th Annual ACM Symposium on Theory of Computing (STOC 2003)*, pp. 116-125, 2003.